# Binary Clustering of Graph Vertices


## ABSTRACT

PURPOSE: Explore the effectiveness of the Fiedler vector for clustering a given data set and a practical introduction to MATLAB

METHODS: The graphs were clustered using binary clustering of vertices, an algorithm which clusters vertices using the Fiedler vector. These were then graphed using a plot function provided by the instructor

RESULTS: On the given dataset there were two clearly defined clusters with little shared edges between the two clusters. We are also able to confirm the edges by confirming the values and graph.

CONCLUSIONS: The binary clustering of graph vertices is an effective method of clustering two groups of vertices. Its effectiveness decreases as the clusters share more edges or in the event of more than two clusters.


## INTRODUCTION

The objective of this report is to explore the effectiveness of using the Fiedler vector for clustering a given dataset. The exercise also doubles as a practical introduction to MATLAB

First named “algebraic connectivity” by Miroslav Fiedler who discovered this method in 1989, we now know this method of clustering vertices as binary clustering of vertices.

The scientific question is how effective is binary clustering of graph vertices. This was tested by looking at the number of edges between the two clusters and comparing it to the total edges in the dataset.

## METHODS

The algorithm used to cluster the vertices using the Fiedler vector first involves forming an adjacency matrix from the given edge list. Then use the adjacency matrix to find the degree

matrix. Use the resulting degree matrix and form the Laplacian matrix by subtracting the adjacency matrix from the degree matrix. Calculate the eigenvalues and eigenvectors using the degree matrix. Obtain the Fiedler vector using the eigenvectors.

Testing was done by comparing the generated graph and sets of clusters using a “testedge” dataset. This would then be compared with the known set of clusters and graphs provided in the assignment documents. This process helps determine any errors in the implemented MATLAB process. Apart from this, console outputs and workspace variables were also used to identify errors in implementation.

The data was randomly generated and then provided by the instructor. We are provided multiple sets of data including other students and the example data but we are expected to only demonstrate the assigned data.

## RESULTS

The tables and figures must have captions that are concise and describe the data without much need to refer to the methods. Numerical results should be presented with only as many significant digits as make sense.

Table 1: Set Of Nodes In Each Cluster

```
Set 1 vertices are:
     1     3     4     6     8    12    13    16    18    19

Set 2 vertices are:
     2     5     7     9    10    11    14    15    17    20
```

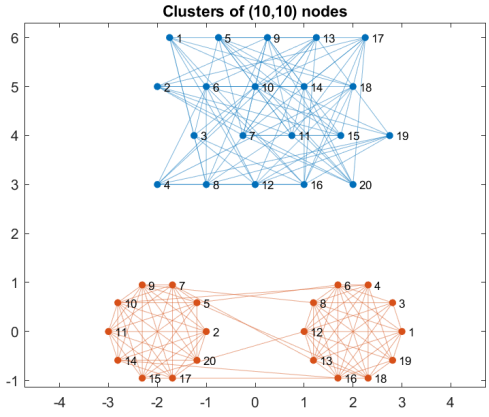


Figure 1: Plot of Clustered Graph

## DISCUSSION

Observations of clustering quality such as number of edges between clusters

- 8 edges connecting the two clusters, when taking into account the total 176 edges. This indicates a good level of separation between the two clusters
- With visual inspection of the created graph we can observe two distinct clusters with little connections. This also indicates a higher clustering quality.

The effectiveness of the Fiedler vector for the given dataset and why it was/was not effective.

- The Fiedler vector was effective for the given dataset. It can be observed from the Fiedler vector that there is a separation between the two clusters with negative and positive values. This property was used to perform the clustering of the vertices.

Do you think the Fiedler vector would be effective at clustering real-world datasets and why/why not?

- Personally, it depends on the dataset, however, I believe that the Fiedler vector would be less effective with real-world datasets. It is clear to me that this method of clustering works better as the separation between clusters is clearer. In real-world situations, data is not always as clearly separated and an increase in connections would decrease the effectiveness.
- This method as its name suggests only works on datasets where the vertices can be clustered into two groups. In real-world datasets, there are often multiple less-defined clusters that this method would be unable to cluster properly.